Statistical Data Editing Design Principles for Household Surveys, With Applications to a Computer Assisted Interviewing (CAI) Environment

Authors: Howard Williams, Elise Kennedy and Dr. Siu-Ming Tam.

Paper presented at The Methodology Advisory Committee meeting, Australian Bureau of Statistics, 14 July 2000.

STATISTICAL DATA EDITING DESIGN PRINCIPLES FOR HOUSEHOLD SURVEYS, WITH APPLICATIONS TO A COMPUTER ASSISTED INTERVIEWING (CAI) ENVIRONMENT

Howard Williams, Elise Kennedy and Dr. Siu-Ming Tam.
Australian Bureau of Statistics.

Abstract

Statistical data editing is redefined as one element of a total quality control strategy for CAI surveys, aimed primarily at preventing non-sampling error, with error mitigation playing an important but subordinate role. Prevention of non-sampling errors is achieved by improving survey processes. This relies on the collection and analysis of data relating to editing performance and to the sources, types and distribution of errors in the data. This type of information should also be used to assess data quality and to provide users with information that can assist them to understand the limitations of the data. In relation to error mitigation, this paper emphasises the need for the systematic and orderly specification of edits and that the amendment of data should occur only in response to important errors. A balance must be achieved between edits applied in the field and those applied in the office, between automated and clerical approaches to verification and amendment of errors, and also between the use of micro and macro-editing methods.

Acknowledgments

The authors would like to thank the following people for their generous assistance: William Winkler and Janis Brown of the US Bureau of the Census; Leopold Granquist and Ake Pettersson of Statistics Sweden; John Kovar, Sylvie Michaud, Barbara Diskin, Philip Giles, Alison Hale and Nancy Darcovich of Statistics Canada; Tony Manners and James Gray of the Office for National Statistics; Mark Pierzchala and Diane Sickles of Westat; and Ton de Waal of Statistics Netherlands and Josée Nollen of Eurostat. While some of the information has been drawn from other statistical agencies and organisations , the authors take responsibility for any errors of fact.

1. Introduction

1. Survey managers currently have access to a larger set of "scientific" techniques than has ever been available in the past to assist with the data editing task. Despite this, statistical data editing remains both an art and a science. This paper argues for a greater focus on the "science" and less on the "art". It is clear that the circumstances under which the methods in the editor's "tool kit" are applied, and the way in which they are combined, needs to be more disciplined and systematic. Often, editing strategies place too much emphasis on "how to edit?" and not enough emphasis on "why edit?" or "what to edit?".

2. The underlying thesis of this paper (and that of the literature) is the need to go beyond the traditional "detect and correct" concept of data editing. A more balanced

and outcomes focussed view of statistical data editing is required. The notion that data editing is solely concerned with error detection and correction is misguided. Errors are symptomatic of the imperfect world in which surveys are conducted. It is not possible to completely eliminate errors, nor is it desirable to attempt this, because in practice this can lead to "over-editing", which has the potential to introduce additional errors. The deployment of too many resources towards error detection and correction also ignores the more powerful and cost effective role that the prevention of error can play in contributing to overall survey quality. Statistical data editing should be considered as one element of a total quality control strategy for surveys.

3. This paper attempts to draw together the theory and practice of statistical data editing in order to define what could be described as "international best practice". The paper also aims to measure the extent to which editing strategies for household surveys in the Australian Bureau of Statistics (ABS) conform to this benchmark. The ultimate objective is to delineate some general design principles and guidelines for the development of editing strategies, which could be applied to ABS household surveys. The pursuit of these goals has been prompted by the conclusion that the potential benefits of a more extensive application of Computer Assisted Interviewing (CAI) to ABS household surveys, could only be realised if the re-engineering of all statistical processes, including data editing, was undertaken. Therefore, in the course of the following discussion, any constraints and opportunities that CAI might bring to the editing task will be identified.

2. The Conventional Definition and Philosophy of Editing and its Limitations

4. Editing is often conventionally described as a process that "cleans up" survey records, that is, it detects and corrects erroneous data. (Granquist and Kovar 1997, Bethlehem and van de Pol 1998, US Federal Committee of Survey Methodology (FCSM) 1990). Records are considered to be "dirty" if data is missing, if there are inconsistencies within a record, if there is invalid information in any field, or if the data recorded appears unusual. "Cleaning up" the data involves editing responses at the record level to provide data which is complete, valid, consistent and plausible. Most editing systems use some sort of flagging arrangement to identify responses which do not comply with predetermined rules (commonly known as "edits"). Those responses "failing" an edit would then be subject to review and possible amendment by editors.

5. Narrowly defining the goal of editing as an attempt to clean up dirty records, combined with the realisation that the use of computers can facilitate and speed this task, has led to the false assumption that the more edit checks that can be performed, the better. It is believed that this will significantly increase the probability that all serious errors will be detected and corrected, and hence the data will be of higher quality (Granquist 1984,1997). Acceptance of this assumption by survey managers has had a number of consequences for current editing practice.

6. Firstly, editing has become a very costly part of the survey cycle. A survey of US federal statistical agencies showed that the median cost for editing demographic surveys comprised 20% of the entire survey budget. The high cost was a major concern for most agencies surveyed (FCSM 1990). Secondly, the same study concluded that "over-editing" occurs in most surveys. Over-editing refers to the fact that beyond a certain point, much of the review and amendment of survey responses is a waste of time and resources. This is because most queried responses are either confirmed or remain unchanged (due to a lack of additional information to make a decision), or if amended, result in changes that can either cancel each other out, or which make no significant contribution to the final outputs of the survey (Granquist and Kovar 1997, van de Pol and Bethlehem 1997, Lepp and Linacre 1993). Excessive editing can result in reduced data quality by introducing additional errors and distortions to the data, and by masking problems in data collection and difficulties that respondents may have with reporting certain information (Granquist and Kovar 1997).   Thirdly, there is a tendency for survey managers to allocate insufficient resources for error prevention and to adequately address the other quality issues peculiar to each survey (Granquist 1984).

2.1 Error Detection

7. Given the traditional perspective on editing, it is not surprising to learn that much of the literature is confined to developing methods that rationalise the process of detecting and correcting erroneous data. For household surveys, a great deal of effort is commonly devoted towards "micro-editing". This process checks individual data records for sequencing errors, consistency and other potential random or systematic errors. Checking for potential errors is achieved by applying formal or heuristic rules to individual data item responses (or codes), or to related data item responses (or codes). These rules are applied within or between records (household, income unit, questionnaire, etc) to flag potential errors. They act as logical constraints on the data (Giles 1988). Where necessary (and possible) the records are "corrected", either clerically, by re-contacting the respondent or by making a judgement based on knowledge of the subject matter and ancillary information, or automatically, by using imputation programs.

8. There are a number of limitations of micro-editing that impinge on its effectiveness and efficiency. Granquist (1990) points out that micro-editing alone cannot always be considered effective as it may not detect even serious errors, for example, unanticipated systematic errors. In relation to efficiency, an obvious shortcoming of micro-editing is evident when query edits are used to flag responses which are potentially incorrect, for example, a range edit applied to quantitative data. In these instances the editor has to manually review all responses flagged as potentially in error, to verify the presence of actual errors. Obviously, it would be much more efficient if the editor could first identify those responses which are more likely to be in error, thus reducing the number of responses that need to be reviewed.

9. In order to make the process of error detection more efficient, a number of other techniques have been developed (mainly for business surveys) as a substitute for, or as a complement to conventional micro-editing. These methods can be grouped

into two broad categories, namely "selective editing" and "macro-editing". They have proved to be more efficient because they considerably reduce the amount of sifting of responses required, without any diminution in the number of important errors detected (Granquist 1997, Bethlehem and van de Pol 1998).

10. Selective editing involves the application of significance, error likelihood criteria or score functions which are used to select and rank those records, which if amended, might have the greatest impact or cause important changes to the estimates. In effect, selective editing techniques employ micro-edit rules with "efficient acceptance limits" (Granquist 1990). This allows the editor to prioritise (in order of error likelihood or impact on estimates, etc) the review of those responses flagged as potential errors (Granquist 1990, Granquist and Kovar 1997). Records are checked in priority order until some predetermined stop value is reached.

11. Macro-editing involves the comparison of aggregates, or the comparison of responses to distributions of variables or observations, to target potential errors in the data. Techniques such as the Aggregation Method and the Distribution Method are included in this category.

12. In the so called Aggregation Method, aggregates from the current cycle of a survey are compared to those of the previous cycle. Those aggregates which have values or counts that deviate substantially from previous periods or from what was expected, are identified. The individual records are then inspected for errors (Bethlehem and van de Pol 1998, Granquist 1994, Granquist 1990).

13. Another very common macro-editing technique is the Distribution Method. This involves forming the raw data for a particular data item, or combination of items into a distribution. The responses for individual records are compared to this distribution. If some records contain extreme responses that appear to be unusual or atypical, then they might require further review and possible amendment (Bethlehem and van de Pol 1998).

14. As for selective editing, macro-editing techniques reduce the amount of manual review required. However, some macro-editing techniques, such as the Aggregation Method described above, have the potential to introduce bias. Bias may occur because any subsequent amendments might tend to be made in the direction of the editor's expectations. Additionally, in household surveys, where individual records are weighted using demographic data, the contribution of each record to the estimates is approximately equal (van de Pol and Bethlehem 1997), which means that selective editing and macro-editing techniques that apply significance criteria will be of limited usefulness in household surveys (Granquist and Kovar 1997).

15. It must also be recognised that, as for micro-editing, macro-editing methods alone cannot always detect systematic errors that were not anticipated when the survey was being developed (Granquist 1990). Sometimes systematic errors are detected by chance as a by-product of the application of edit rules. For example, it may be noticed that the majority of failures for a particular edit were due to non-response, and this may indicate that many respondents were not able to fully understand a question (Granquist 1984). A more rational approach would involve

the development of tools to detect and gather intelligence about these types of systematic errors, so that they could be avoided in future iterations of a survey, or at least their nature and impact on the data could be notified to users (Granquist 1984).

2.2 Error Correction - The Fellegi and Holt Approach

16. Error correction has traditionally been mostly a manual process. The use of CAI for household surveys permits some edit failures to be corrected in the field by interviewers conferring directly with respondents. However, the majority of the more complex edits are applied in-office where the option of re-contacting respondents is not practicable. Batch oriented in-office editing programs usually provide consolidated edit listings and often they show that a record has failed a number of edits. The problem is then to determine for each record if the edit failures have been triggered by a series of errors, by a single error, or by multiple errors, each of which has resulted in a string of consequential edit failures. This is known as the "localisation problem". Straightforward random errors can be dealt with easily, but sometimes in order to "correct" the data, there is a need for the editors, using their subject matter expertise, to make a judgement call based on other corroborating evidence elsewhere in the record. This procedure amounts to quasi-imputation and it is a laborious and time consuming task. A number of passes through the editing system may be required, because the editor may not necessarily know if the changes made will cause the record to violate other edits. Additionally, the application of formal or heuristic rules used for clerical amendment of responses are prone to inconsistent application by processing staff. It is also possible that large numbers of amendments may lead to the data being distorted, in respect of both the quantum of the estimates, and the univariate and multivariate distributions (FCSM 1990).

17. In what is generally acknowledged as a landmark paper, Fellegi and Holt (1976) describe a methodology for amending erroneous responses which avoids the problems associated with manual methods. The methodology is particularly useful in circumstances where consequential edit failures have been triggered, and it is not immediately obvious to the editor which field is at fault. In the Fellegi and Holt approach, the basic edit rules are specified by the relevant subject matter experts. By converting these rules into the "normal" form, logically "implied" edit rules can be derived. Together, the specified and implied rules form a "complete" set, and from this set it is possible to identify the minimum number of fields that require amendment, and to automatically deduce the range of responses that satisfy all related edit rules. Importantly, the amended responses deduced in this way will not result in distortions of the univariate and multivariate data distributions. This approach also ensures that amendment rules are uniformly applied and records will require only one pass through the editing system to make them "clean".

3. Editing as a Network of Interdependent Quality Control Processes

18. The ultimate desired outcome of any editing strategy is data of improved quality (see Fig 1). Better data quality relies on the continual improvement of survey

processes in order to prevent errors occurring, and on the mitigation of those errors that could not be avoided. As a by product of the editing strategy, an assessment of the quality of the data should also result. This assessment can provide a benchmark, against which overall improvements to data quality can be measured, as each iteration of a survey is completed. It can also assist the users of the survey output to better understand the limitations of the data, especially if they intend to undertake further analysis.

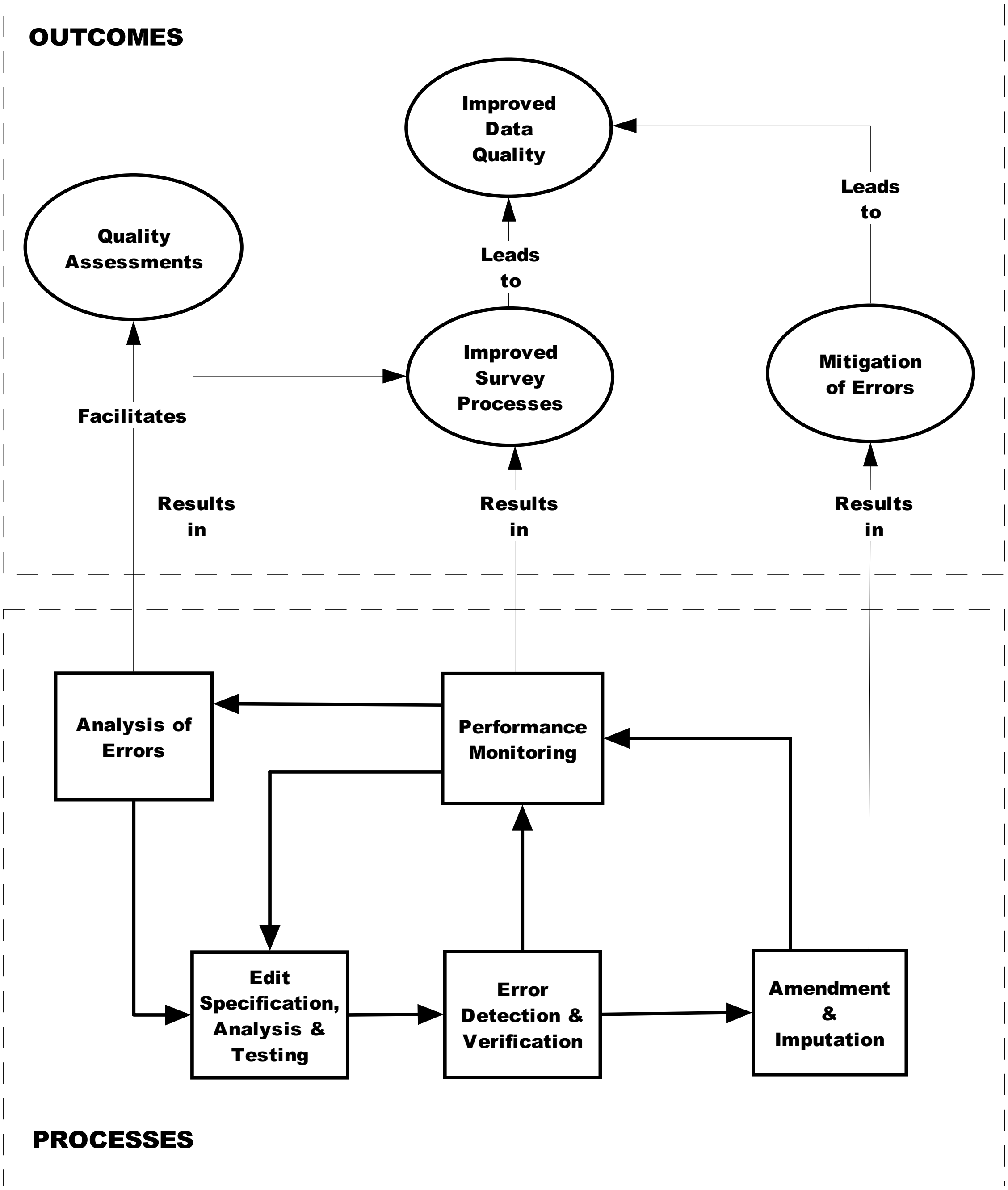

Fig 1. - Editing as a Network of Interdependent Quality Control Processes

19. As stated previously, the conventional definition of editing is not particularly helpful because it focusses on the cleaning of "dirty" data. This can lead to over-editing and have other negative consequences on data quality. Granquist (1984, 1997) and Granquist and Kovar (1997), consider it preferable to extend the traditional "detect and correct" concept of editing, to include processes which provide information that could be used to assess the quality of the data, and for improving other phases of the survey cycle so that potential errors are avoided. They maintain that these additional processes should be the primary (but not exclusive) focus of editing. The literature also indicates that there is a need to review editing practices in the light of Granquist's perspective, and for the additional processes that his view implies to be considered as an important component of statistical data editing (FCSM 1990, Lepp and Linacre 1993, Bethlehem and van de Pol 1998).

20. Given the constraints imposed by finite survey resources, and the acknowledged limitations of editing methodologies, it is not possible to anticipate or locate all errors, and not desirable in practice to "correct" all errors so detected. Under these circumstances, and following the lead provided by Chinnappa, et al (1990), Granquist and Kovar (1997) and Bethlehem and van de Pol (1998), statistical data editing, can be described as:

> "A network of interdependent quality control processes, organised to identify and help prevent potential non-sampling errors, and to mitigate and then assess the impact of known non-sampling errors on the quality of survey outputs."

The primary focus of editing should be on the identification and prevention of potential errors. This is best achieved by gathering and analysing intelligence in order to find and investigate significant patterns in the survey data which may be indicative of unanticipated systematic errors. This information can then be used to improve survey practices.

21. A secondary aim is to mitigate the impact of known errors on survey quality. Error mitigation should be focussed on important errors and achieved by applying rules and methods for error detection, amendment and imputation, that are appropriate, effective, efficient and statistically defensible.   "Important errors" are those that would be obvious to the editor and to users, and which would reduce the credibility of the estimates and / or would cause later problems for down stream processing. "Appropriate" means that the right type of rules or methods are applied, having regard to all relevant circumstances impinging on each case. For example, a range edit should only be applied in situations where it is possible to develop limits on some objective basis, drawn from knowledge of the subject matter, and not on some prejudiced view of the expected distribution of responses. "Effective" means that all the important errors are detected and amended or imputed if possible. An "efficient" edit is one for which relatively few false alarms are triggered. From a practical perspective, an "efficient" error amendment or imputation method is one that does not cause consequential edit failures. "Statistically defensible" means that the amendments to data arising from correction and imputation of errors should, on the whole, be repeatable and not cause the data to change in unexpected

directions. It is especially important that the underlying univariate or multivariate distributions should not be distorted (Barcaroli and Venturi 1997).

22. Management information relating to the performance of editing is required to ensure that rules and methods applied are effective, efficient and statistically defensible. The end result should be improved survey practices and survey data having a level of consistency, integrity and coherence sufficient to allow publication (Granquist and Kovar 1997).

23. The core processes involved in this editing network can be broadly described as: analysis of errors; edit (rule) specification, analysis and testing; error detection and verification; error amendment and imputation; and performance monitoring (see Fig 1.). Keeping in mind the desired outcomes, the network of processes is organised in the form of a feedback loop. Each of the processes and their interrelationships are described in more detail below.

3.1 Analysis of Errors

24. This process should identify the major sources of non-sampling error pertinent to each survey. This information can be used to help with the specification of edits and to improve other survey processes, so that errors can be prevented. Error analysis also provides information to assess the overall quality of the survey data. A quality assessment provides a benchmark, against which the success or otherwise of an editing strategy can be measured. Quality assessments can also assist the users of the survey output to better understand the limitations of the data, especially if further analysis is required.

25. An analysis of errors may be conducted at various points throughout the survey cycle, depending on the individual circumstances of each collection. Ideally, it would be preferable to identify the sources of potential errors in sufficient time to allow the survey methodology and procedures to be changed, to avoid the occurrence of these errors. This is possible for on-going surveys, because information from the previous cycle would be available for analysis during the development phase of the current cycle of the survey. However, for adhoc surveys, a complete analysis of errors is only possible at the end of the cycle.

26. Van de Pol and Bethlehem (1997) outline an extended version of an error classification devised by Kish (1967). They identify five sources of non-sampling error. The first three sources are overcoverage, measurement and processing error which arise during collection, data capture and processing. The remaining two sources are undercoverage and non-response error and they occur because no survey measurement was possible. However, from an editing perspective, a more crucial classification is one based on the type of errors and their impact on survey quality. With this in mind, it is better to differentiate between random errors and systematic errors.

27. Random errors, as the label implies, are errors that exist randomly throughout the data that do not have a tendency to consistently change the estimates in any one direction. They primarily arise due to in-attention by respondents, interviewers

and other processing staff during the various phases of the survey cycle. Some random errors will be immediately obvious to the editor and to users, but others will be buried in the matrix of data and may never be detected. Only a very small proportion of detectable random errors are important, that is, few have a significant impact on the quality of survey outputs or cause later problems during processing (Granquist 1984).

28. Systematic errors are those that arise because respondents or processing staff do not understand or misinterpret concepts, definitions or the questions being asked, or because of faults in the conceptual or procedural elements of data collection, capture and processing. They may also be due to the respondents' deliberate action (motivated by malice, the sensitivity of the question, etc.) to refuse to answer questions, or to provide plausible but nevertheless spurious, misleading or incomplete responses. Systematic errors have far greater potential to affect the quality of survey outputs than random errors, because, if for example, large numbers of respondents misinterpret a question in the same way, then a bias may be introduced into the estimates.

29. Some systematic errors can be anticipated, as knowledge of their existence may come from field instrument testing and the like, and many of these may be prevented by appropriate remedial action. Others may become evident once a consistent pattern of errors has been identified, either during the processing of the survey (via the performance monitoring of edits), or later when outputs are analysed. In this regard, analysis of unedited responses to detect unusual patterns in the data, may indicate problems with particular questions or interviewer bias (Fries and Woodburn, 1995). For these, remedial action might only be possible for the next iteration of the survey, for example, systematic errors arising from poor CAI instrument design.

3.2 Edit Specification, Analysis and Testing

30. The rational specification of edits is based on four sources of information. The first is knowledge of the key data items, how they might be used by survey clients and the level of quality required. Quality includes accuracy, consistency and timeliness. In relation to non-sampling error, accuracy is best expressed in the form of data item priorities and boundaries for the total error of the estimates, eg. "estimates for data item X should lie within plus or minus 3% of the true value" (ABS Editing Manual 1993). Previous or similar surveys (questionnaires, user requirement surveys, pilot test and dress rehearsals, etc) are a convenient source of information but there is no substitute for a detailed specification of survey output requirements developed with the major clients.

31. The second source of information arises from an understanding of the social and economic conditions that are likely to influence respondents and the implications they have for the relationship between data items (ABS Editing Manual 1993). In this regard, subject matter knowledge, as well as applying methods such as principle components analysis, correlations, graphical analysis and other exploratory data techniques, is useful. Together, these techniques make it possible to describe the distribution of important variables, group similar variables together,

discover relationships between variables and useful ratios, all of which can greatly facilitate the specification of appropriate edits (Whitridge and Kovar 1990).

32. The third information source is a knowledge of the type and potential impact of errors on quality obtained from a thorough analysis of errors. The focus should not be to develop rules to cover all possible checks, but rather to develop edits to detect serious errors (Granquist and Kovar 1997). With this approach rules can be targeted at the important random errors, and at the anticipated systematic errors, for which no preventative action could be devised prior to the field phase. Confining the scope of edit rules in this way will help to prevent over editing.

33. The objective of using information from these three sources is to specify the minimal set of rules required to define the acceptance region for an item or related data items. This is better than defining a rejection region, because then there is no danger that an unforeseen response will automatically be assumed as acceptable. After initial specification, it is necessary to analyse the acceptance regions of each set of rules to eliminate conflicts or redundancies (Whitridge and Kovar 1990). The derivation of "implied" edits, as proposed in the Fellegi and Holt approach, can be of great assistance in this process. They can indicate relationships between the original edits that the subject matter expert may consider to be invalid, and therefore changes to the initial set will be required (Greenberg 1988). Extra care is needed to ensure that there are no conflicts between edits that might be coded into the CAI instrument and those that are to be applied later in the office (Dufour, Kaushal, and Michaud 1997).

34. However, although all such analyses described above are necessary, they are not sufficient, and a fourth source of information should be tapped. Information about the performance or likely performance of edits is also needed, in order to ensure that they are effective (detect all important errors) and efficient, that is, that they target those responses that are most likely to be an error and therefore do not generate too many false alarms. Pilot testing of each edit, and the set of edits as a whole, is required to observe their likely net effect, and to ensure that over-editing does not occur. Performance monitoring of edit effectiveness and efficiency during the processing of a survey can also periodically be fed back, to enable for example, real-time adjustment of edit tolerances. CAI survey instruments are particularly suitable for gathering and feeding back this sort of management information. Vigilance is required when analysing the results of performance data to be certain that the univariate and multivariate distributions have not been distorted by amendment of edit failing responses (Barcaroli and Venturi 1997).

3.3 Error Detection and Verification

35. Often users of social and socio-economic data request that the collection agency provide unit record files to permit detailed investigation. In these circumstances the agency may feel obliged to eliminate all sequencing errors and inconsistent responses in the data file. For household surveys using traditional pen and paper (PAPI) data collection methods, most editing is directed towards identifying questionnaire sequencing errors and inconsistent responses. The advent of CAI has largely eliminated the occurrence of sequencing errors and has permitted

interviewers to query many inconsistent and implausible responses in the field. However, the focus must remain on detecting and verifying only the important errors, and to this end, a balance needs to be struck between micro-editing in the field and micro-editing in the office, and also between micro-editing and selective and macro-editing methods.

36. Ideally, as many edits as practicable should be included in the field instrument, where it is most likely that "correction" of errors can be effected. The constraints include the performance limitations of the laptop computer, length of the interview, respondent load, the ability of the interviewers to resolve the edits, the sensitivity of the questions and the likely impact of follow-up questions on response. As a minimum, the following should be included: sequencing edits; checks that identify households for which no further processing will occur, eg, out of scopes, partially responding households ineligible for imputation, etc; edits that check responses which are used to produce key estimates; and edits that check variables used for post stratification (Dufour, Kaushal, and Michaud 1997).

37. Wherever possible, selective and macro-editing techniques should be applied because they provide a more efficient method of detecting important errors. For example, using selective editing, it is possible to stream edit failures into two categories. One category will include those edit failures that are more likely to contain serious errors and which should be subject to manual verification. The remaining edit failures in the second category, once verified as errors, could be automatically amended (Granquist and Kovar 1997), perhaps using the Fellegi and Holt approach or similar, thus avoiding the expensive process of manual error localisation. Granquist and Kovar (1997) support their argument by referring to studies of business surveys employing the traditional micro-editing approach. The results of editing evaluations in these studies indicate that all the important errors would have been detected if only 5 to 10% of the amended responses had been flagged for manual review.

3.4 Error Amendment and Imputation

38. Error amendment remains a predominantly manual process in many statistical agencies, mainly because there is a need to apply the subject matter expertise of the editor to verify and localise the error. Nevertheless, amendments to responses found to contain errors often amount to little more than quasi-imputation and this can have negative consequences for data quality as noted in section 2.2. The Fellegi and Holt method can automate the amendment process in a way that avoids these limitations. That is not to say that all amendments should be derived automatically. A balance between manual and automated methods needs to be considered in the light of resource constraints, the sophistication of any automated system proposed, and a realistic assessment of how much error mitigation is required to make the data suitable for publication (Granquist and Kovar 1997).

39. It is not possible to be prescriptive about the type of imputation that should be used for household surveys generally, because there are many considerations that need to be accounted for, and these vary enormously depending on the topic under investigation. However, the point in relation to data editing is the requirement to

select a procedure that can be integrated with the editing strategy, so that any imputed values will not violate other edit rules (Greenberg 1986). In the Fellegi and Holt method, any fields that have been imputed will always satisfy the edit constraints. In doing so, the procedure takes into account all the other original responses in the record that are logically related to the imputed field (Fellegi and Holt 1976).

3.5 Performance Monitoring

40. Essentially, performance monitoring involves the collection and analysis of a range of summary statistics and diagnostics, which are compiled at different levels of aggregation and at different points in time during the survey cycle. Depending on the operational details of each survey, the data might relate to the survey as a whole, to groups of records or individual records, or to sets of edits or each individual edit. The data could include cumulative record counts, data file snapshots, record audit logs, coding and edit logs, etc. For example, Engstrom and Jansson (1996) suggest some specific numerical measures which include measures of the proportion of records failing edits and the proportion of failed records which are changed due to the editing process. They also suggest use of indicators generated for each specific edit.

41. The monitoring of editing performance is a vital component of any high quality editing strategy. Monitoring provides an important source of intelligence to assist specification of effective and efficient edits, the fine tuning of edit tolerances, the measurement of the impact of amendment and imputation on survey estimates and costs, and it can also provide important information to facilitate error analysis and to assess the quality of responses. In many cases, the authors consider that survey designers do not give sufficient consideration to compiling performance data and its accessibility for analysis. Sometimes the volume of data is overwhelming, but more often too little data is available. Equally, even when the data is sufficient and accessible, survey managers tend to be disinclined to analyse the collected data in a systematic way to derive the maximum benefit. It is also not uncommon to find that the compilation and analysis of editing performance data has not been considered at all in the design of some edit strategies.

4. Practice in Selected Statistical Agencies

42. A number of statistical agencies and organisations were contacted to obtain additional information about current editing practices for their household surveys. The agencies and organisations that responded were the Office for National Statistics (ONS) in the UK, Statistics Canada, the US Bureau of the Census, Statistics Sweden, Statistics Netherlands and Westat in the US. The following is a brief summary of their experience of statistical data editing.

4.1 Editing Strategy Guidelines

43. Most agencies do not have a generic set of guidelines to assist with the development of editing strategies but instead rely on the application of internal

standards. Often the justification for the lack of this documentation is that edits are written specifically for each survey, so only a few standard edits exist. The US Bureau of the Census and Statistics Netherlands have no guidelines or documentation for population or household surveys. They do, however, consult with clients to develop some general strategies and specific guidelines for each survey. Westat also find it is difficult to establish a standard methodology for editing practices because their operating environment is different to official statistical agencies, as they work on a contract basis and the clients usually have specific editing requirements. The ONS have a standards and quality assurance team which monitors editing practices, and the standards they administer are embodied in standard blocks of (BLAISE) code. In contrast, Statistics Canada apply a very comprehensive set of editing guidelines called the Statistics Canada Quality Guidelines (Statistics Canada 1998).

4.2 Editing and CAI

44. In the larger agencies, CAI is the preferred method of data collection and capture for most of their household surveys. There are differences between agencies in the balance of motives that have prompted the use of CAI. Nevertheless, it is clear that the underlying expectation is that CAI will deliver improvements in the efficiency of data collection and capture and in the quality of input data. In theory, CAI instruments can improve the quality of responses by automating the sequencing of respondents through the schedule of questions; by allowing edits to be applied and resolved in the field with the assistance of the respondents; by gathering both quantitative and qualitative information about initial responses and amendments and about the performance of the edits themselves. However, in practice there are constraints that limit the extent of the quality gains.

45. Most agencies have found that routing and skip pattern errors have been reduced but not entirely eliminated by using CAI. For example, the US Bureau of the Census find that sometimes off-path information still occurs because their software is not able to adjust to situations where the interviewer is required to backtrack to alter responses to previous questions (Bowie et al 1998).

46. Statistics Canada have discovered that respondent load is often too large if all possible edits are included. In ONS, their experience is that for a number of surveys, some edits are far too complex to apply in the field, because respondents are not able to provide additional information to resolve the query. But for other surveys, including their Labour Force Survey, no further editing is required after the field phase. For these and similar reasons, all agencies limit the number of edits applied in the field and choose to supplement them with additional checks in the office, which can take the form of micro-edits or macro-edits, or both. With the exception of ONS, the agencies contacted report that this in-office editing includes re-application of some of the micro-edits that were included in the CAI instrument. This occurs because some of the query edits are used to filter responses in the field and are re-applied as fatal edits in the office. For example, in the Canadian Labour Force Survey, the data from the CAI instrument are edited a second time using the rules in the CAI instrument and then a more complex set of edits are applied (Rowland 1994).

47. Whilst CAI has not eliminated the necessity for some post-field editing, other benefits of CAI have been realised. Most of the agencies contacted believe they are receiving better quality data than they would normally expect to get using PAPI collection methods. In the experience of Statistics Canada, CAI has significantly reduced edits dealing with "consistency of flows". This allows a greater focus on editing dealing with consistency between fields. The ONS has also experienced a reduction in the cost of editing. In this way many benefits have been realised through CAI.

4.3 Intelligence Gathering and Performance Monitoring

48. In large statistical agencies, the gathering of intelligence relating to the sources and types of errors and the monitoring of editing performance is traditionally collected by way of editing reviews. They are a crucial part of good editing strategies, but in practice reviews are costly and time consuming to conduct. The US Bureau of the Census do not conduct editing reviews. Statistics Netherlands review their editing methods and practices on an adhoc basis for some individual surveys. The ONS conduct internal reviews of their edit processes as part of a general quality assessment prepared for each survey.

49. Statistics Canada (1998) conduct thorough reviews of editing for all their household surveys. The reviews include complete lists of edits detailing individual justifications for the application of each edit. Reports are often published that indicate what the estimates would have been if edits had not been applied, sometimes together with a description of how records were edited (Fournier 1997). Reviews are very comprehensive and attempt to answer questions such as, is a response failing the edit because of the assumptions underlying the edit? or is the response failing the edit because the respondent has misunderstood the concept in the question? (Hunter and Ladds 1995) In this way, Statistics Canada use their edit reviews to provide information to questionnaire designers for future improvement of the survey instrument. Reviews are also used to inform interviewers of problems with previous iterations of the survey (Statistics Canada 1999).

4.4 Use of Editing Packages

50. The agencies contacted report the use of two types of software for statistical data editing. In general, agencies apply some basic edits as part of the collection instrument, then as noted above, a second set of edits is applied when the data arrives in the office. For many of the agencies this gives rise to two sets of software being used during the editing process.

51. BLAISE is commonly used for data collection and capture in CAI instruments. Currently, Statistics Netherlands, the ONS and Westat use this package and the US Bureau of the Census are presently developing a data collection system based on BLAISE. The other agencies contacted did not state specifically which software they used in their data collection instruments.

52. In-office editing tends to be conducted by each agency using various combinations of proprietary and / or other tailored software developed in-house specifically for this purpose. Statistics Canada use SAS code for instances where consequential edit failures have not occurred and therefore no error localisation is needed. In circumstances for which it is not clear where the error lies, and localisation is therefore required, Statistics Canada have developed a software package called the generalised edit and imputation system (GEIS) which applies the Fellegi and Holt approach. The ONS uses BLAISE to impute for inconsistent information. For macro-editing the ONS use SPSS. Westat uses a system written in SAS. The US Bureau of the Census has no generalised editing system at present and uses programs written in SAS or Fortran.

53. Quite a few countries either have or are developing generalised editing systems. The US Bureau of the Census are investigating the possibility of using a generalised editing system. They are considering the DISCRETE edit system for person or household surveys or AGGIES.   They may also investigate using CANCEIS, a package Statistics Canada use for their Census. All these systems attempt to generalise the editing process in some way. There is a question as to whether generalised systems should be used or if instead it would be better to create editing systems tailored to each survey. Some survey managers believe that because each survey is different, a flexible approach to the content of editing systems should be employed to cater for these differences. However, the coding of editing systems designed individually for surveys is time consuming, expensive and error prone. Others suggest that consistency across surveys would reduce confusion and errors and therefore, creating reusable software modules which read and use edit rules has been suggested as the best approach (Kovar and Winkler 1996).

4.5 Imputation

54. Statistics Canada, the US Bureau of the Census and the US National Agricultural Statistical Services (NASS) have adopted the Fellegi and Holt methodology for their imputation systems. Statistics Canada have developed CANEDIT and GEIS, the US Bureau of the Census, SPEER and DISCRETE and NASS, AGGIES. These systems determine the minimum number of variables to impute, then perform the imputation, often using hot deck imputation (Bankier 1999). According to de Waal (2000) there are not many systems based solely on the Fellegi and Holt methodology because it is hard to develop the operations research techniques needed for the model. One problem with the Fellegi and Holt approach is that edit and imputation problems, in certain circumstances, can be very large and complex. In these situations, some agencies have encountered difficulties implementing the methodology because of the practical limitations. Statistics Canada have therefore developed another edit and imputation system called CANCEIS. In contrast to the Fellegi and Holt approach, this system first searches for donors, then performs imputation by finding the minimum number of changes given the donors. It is capable of performing imputation for both qualitative and numeric data simultaneously. This results in a much quicker and more efficient system, it can solve larger edit and imputation problems than systems which follow the standard Fellegi and Holt methodology (Bankier 1999).

5. Editing Practice in the Australian Bureau of Statistics (ABS)

55. The ABS has a very extensive program of household surveys, covering a wide range of social and socio-economic data. For this paper, five surveys encompassing the range of data typical of the program were reviewed, to provide a general picture of the current nature of editing practice for household surveys in the ABS. The surveys reviewed included the National Health and Nutrition Survey (NHNS), the Survey of Disability, Ageing and Carers (SDAC) and the Household Expenditure Survey (HES), each of which are conducted approximately every five years, and also the monthly Labour Force Survey (LFS) and the monthly Survey of Income and Housing Costs (SIHC). SDAC and HES are the only surveys that have been enumerated using CAI based collection methodologies. All of the following observations relate to these five surveys unless stated otherwise.

5.1 Editing and CAI

56. The ABS has been motivated to use CAI more extensively for its household surveys for reasons similar to those that have prompted other agencies to adopt the methodology. Likewise, the same constraints have been encountered. As for other agencies, the number and type of edits applied in the field is limited, and the same field edits, together with supplementary input edits are applied later in the office. The more complex edits are applied during output processing of the data. Off-path information does sometimes occur in situations where the interviewer has been required to alter responses to questions asked earlier in the interview, and this has to be removed during later editing. A more detailed outline of editing practice in the ABS follows.

57. Editing of household surveys data occurs in two distinct phases, the input phase and the output phase, and this approach applies to both PAPI and CAI based surveys. For CAI based surveys, basic input edits (consistency and logical edits) and some simple query edits are coded into the instrument and applied in the field. Data from completed interviews is transferred from the laptops and passed through an office input editing system, which applies the same edits as were used in the field, as a check that field procedures have been followed correctly. Field operations staff resolve any in-office edit failures and then perform various types of coding, for example, family coding and employment coding. At the final stage of input processing, a "clean" input data file emerges, as all records which had edit failures that could not be resolved (a rare occurrence) are discarded. The input file is passed on to the relevant subject matter area for further output editing, validation and analysis. It is unusual for a respondent to be recontacted after the input file has been delivered.

58. The basic field input edits are specified by field operations staff, in conjunction with subject matter experts. The field query edits, however, are mostly specified by questionnaire design specialists, as it is they who are responsible for developing specifications for the CAI survey instruments. The actual BLAISE coding of the edits in the instrument is completed by field operations staff. The output edits are

specified by the relevant subject matter area. Coding in SAS and application of the output edits to the input data file, is the responsibility of output processing staff. Subject matter experts review the output edit failures and provide amendments, which are subsequently applied to the output file by output processing staff.

59. The general approach to editing is characterised by a heavy reliance on micro-editing, and a belief in the veracity of what Granquist (1984) refers to as the "mass checks approach". This strategy can be directly attributed to the fact that many survey managers' opinions have been strongly influenced by the requirement to provide unit record files to clients for further analysis. The pervasiveness of this philosophy is evident in all of the surveys reviewed, with the exception of the LFS which does not publish a unit record file.

60. There is no standardised approach to edit specification by subject matter areas. During the rule development process, the edits are not formally nor rigorously reviewed, and there is evidence of the use of inappropriate edits and methods, and redundant and inefficient edits (Williams 1999). Subject matter staff, generally without the benefit of adequate data to undertake this exercise in a systematic way, come to a "seat of the pants" decision, and in many cases the majority of rules tend to be mainly historical artefacts. Sometimes this means that the edit specification process continues up until a deadline is reached.   There are even examples of edits being specified and used to resolve serious errors after the release of unit record files to clients.   As a result, the edit specification process is inconsistent between surveys, even surveys conducted by the same subject matter areas. This unstructured approach to edit specification, in which there is a disturbing lack of priority and uniformity, is common throughout ABS household surveys.

5.2 Editing Strategy Guidelines

61. A data editing manual for use with ABS surveys has been available in the ABS since 1993. It acts as a central repository of editing techniques and provides standard guidelines for the design of new editing strategies and the review of existing ones. It is very comprehensive and places editing in the context of the broad framework of statistical practice adopted throughout the ABS (ABS Editing Manual 1993).

62. Much of the content is highly relevant to the editing requirements for household surveys. Nevertheless, it needs to be extended, to provide guidelines which specifically address the unique problems encountered in household surveys, and those surveys using CAI methodologies. Broadly speaking, the underlying approach suggested in the manual accords well with the approach to editing suggested in section 3, although the emphasis remains on error correction rather than prevention.

63. Unfortunately, much of the advice, at least in relation to ABS household surveys, does not seem to have been taken on board and used in practice. In fact, many household survey managers in the ABS are either unaware of the manual's existence, or if they know it exists, are not aware of its relevance to their surveys.

5.3 Intelligence Gathering and Performance Monitoring

64. Editing reviews of household surveys are for the most part conducted on an adhoc basis, when it is evident that there are problems. The review process in the ABS is hampered by a general lack of adequate documentation of editing procedures, insufficient error analysis and performance data, and the by the inaccessibility of whatever data is available. For example, a formal review of editing practices for the SIHC (Williams 1999) required the manual compilation of editing performance data over a period of several weeks. The authors are not aware of any ABS household surveys for which a rigorous analysis of errors is routinely conducted.

5.4 Use of Editing Packages

65. The ABS uses BLAISE to code the field instruments and the associated input edits for CAI based household surveys. Output edits are coded using SAS within the Social Surveys Output System (SSOS) or the Household Surveys System (HSS). Currently, no generalised editing package is used, although a new Household Surveys Facilities (HSF) system is being developed, which will provide generalised input and output processing facilities, including functionality for editing, amendment and imputation.

5.5 Imputation

66. As for other agencies, imputation methods in the ABS are survey specific. In general, if imputation is deemed necessary, strategies employing some form of donor imputation are invariably preferred for household surveys. Inevitably, these methods result in consequential sequencing and consistency edit failures, which are reviewed and repaired as required.

6. Data Editing Design Principles

67. It is evident from the forgoing that a number of clear principles emerge that can be applied to direct the development of editing strategies for CAI based household surveys. These principles are:

> 1. The primary focus of any editing strategy should be on the identification and prevention of potential errors, in particular, on ensuring that anticipated systematic errors are countered by appropriate survey methods and practices, and that facilities are in place to identify unexpected systematic errors.
>
> 2. The mitigation of known errors should be a secondary aim of the editing strategy. Effort should be focussed on the important errors rather than on an attempt to eliminate all errors. This means concentrating on detecting and amending only those errors that will be obvious to editors and to clients (thereby reducing the credibility of the estimates) and / or would cause later problems for down stream processing.

3. The editing strategy must provide facilities to collect and analyse data relating to the sources, types and distribution of errors.

4. The editing strategy must allow for the monitoring and analysis of editing performance data. Specifically, an audit trail must be established and also facilities to automatically produce edit performance diagnostics and summary statistics.

5. Design processes that allow the systematic and orderly specification of edits. The processes must utilise knowledge of the relationships between key variables, user needs, anticipated systematic errors, and if available, edit performance data. The generation of implied edits using a Fellegi and Holt approach or some other method should be incorporated as an important component of these processes.

6. Wherever technically feasible, attempt to design efficient edits that use selective editing and / or macro-editing techniques.

7. Attempt as much editing and coding as practicable using the collection instrument, to realise the maximum benefit from CAI.

8. Strike a balance between clerical and automated error verification and amendment. The less important errors should be automatically amended using a Fellegi and Holt or similar type approach.

9. Provide information to users regarding the quality of the reported data, and the methods used to detect and amend errors and to impute data. Include information about the types of important errors that were amended, and about those errors that remain in the data.

7. Implications for ABS Editing Practice

68. Given that the pervasive editing philosophy applied to household surveys in the ABS tends to conform to Granquist's (1984) description of a "mass checks approach", then considerable effort will be needed to initiate a cultural change, to ensure that the focus of survey managers is redirected towards the design of editing strategies that emphasise prevention rather than "cure". There are a number of ways of initiating this type of cultural change, but fundamental to all of the options is a requirement for education of survey managers and editing staff. Education will only succeed if staff are convinced of the need for a change in their approach to statistical data editing, and this can only be achieved if hard evidence of the consequences of current practice can be brought to their attention. Therefore, the provision of appropriate, timely and accessible editing performance data (including an audit trail) and data about the sources, types and distribution of errors, must be a mandatory component of the final version of the HSF system currently being developed. This type of data can also play a role in helping to assess if the change

in attitude has been applied in practice and has resulted in the outcomes that are expected.

69. Even if survey managers and editing staff are convinced of the need to alter their approach, inevitably some resistance will be encountered. To overcome this, formal management structures will need to be devised to manage the change. This is particularly important in relation to editing practice, because some of the expected outcomes are subtle and not easily verifiable solely by reference to editing performance data and the like, as referred to in the previous paragraph. For example, in relation to edit specification, how do we determine if the proposed edits are appropriate, consistent, and are targeted at the relationships between key data items? The authors suggest that some form of review process be undertaken at key points in the editing cycle to assess that these less visible outcomes have been achieved. The authors also suggest that the editing strategy for each survey should be outlined in broad terms and included in the business plan, for approval by the survey steering committee. The more detailed elements of the editing strategy, including the post-field edits, should be developed in conjunction with the CAI collection instrument.

70. Additionally, assistance needs to be provided to survey managers so that changes to editing practices are made in a systematic, orderly, consistent and statistically sound way. The ABS Editing Manual (1993) is an invaluable tool, but it needs to be updated to include more extensive guidelines which address the unique characteristics of CAI based surveys and household surveys in general. In particular, more examples of selective editing and macro-editing techniques suitable for use with categorical data should be included. More promotion of the manual as a reference source for survey managers to assist in the review of editing strategies would also be beneficial. ABS methodology experts also need to be more proactive and become involved in the development of editing strategies from the very earliest stages. This is vital, particularly if, as the authors suggest, wider use is made of the more sophisticated selective editing and macro-editing techniques.

71. The compilation and provision to users of information relating to the quality of survey data, and the methods used to validate it, will require considerable planning and development. This process must be undertaken cooperatively with the clients to ensure that their needs are met. For example, documentation provided with unit record files must indicate to users the level of analysis for which the data can be expected to be accurate.

References

Australian Bureau of Statistics (ABS) Editing Manual. (1993). Australian Bureau of Statistics, unpublished document.

Bankier, M. (1999). Experience with The New Imputation Methodology used in the 1996 Canadian Census with Extensions for Future Censuses. Working Paper No. 24, Presented at UN/ECE Work Session on Statistical Data Editing, Rome: Italy, June 2-4, 1999.

Barcaroli, G. and Venturi, M. (1997) DAISY (design Analysis and Imputation System): structure, methodology and first applications. Statistical Data Editing, Volume 2 (1997). Geneva: United Nations.

Bethlehem, Jelke and van de Pol, Frank. (1998) The Future of Data editing. Computer Assisted Survey Information Collection. Ed. Lyberg et al. New York: Wiley and Sons.

Bowie, C. (Chief Editor.) (1998). Automated Data Collection How Have the Demographic Surveys Fared Thus Far? Washington, DC: US Bureau of the Census.

Chinnappa, N. Collins, R., Gosselin, J.F., Murray, T.S. and Simard, C. (1990). Macro-Editing at Statistics Canada. Statistics Canada, unpublished document.

de Waal, T. (2000). New Developments in Automatic Edit and Imputation at Statistics Netherlands. Statistics Netherlands, Voorburg.

Dufour, J., Kaushal, R. & Michaud, S. (1997). Computer Assisted Interviewing in a Decentralised Environment: The case of Household Surveys at Statistics Canada. Survey Methodology, Vol 23, Ottawa: Canada, 147-156.

Engstrom, Per & Jansson, Charlotte, (1996) What do we want to know about the Editing Process in a Survey? Proceedings of the Second International Conference on Methodological Issues in Official Statistics; Stockholm, September 1996, Statistics Sweden, 169-172.

Fellegi, I.P. & Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, Vol. 71 Washington: American Statistical Association, 17-35

Fournier, Elaine. Leesti, Tracey and Lauigne, Mylene. (1997). Survey of Income and Labour Dynamics: Processing Strategy for Wave 1 Income Data, Ottawa: Statistics Canada.

Fries, Gerhard and Woodburn, R. Louise (1995) Using Graphical Analysis to Improve all Aspects of the Survey of Consumer Finances. American Statistical Association, 1995 Proceedings of the Section on Survey Research Methods Volume 2, Alexandria: American Statistical Association, 927-932.

Giles, P. (1988). A Model for Generalised Edit and Imputation of Survey Data (GEIS). Canadian Journal of Statistics 1988 vol 16, 57-73.

Granquist, L. (1982). On Generalised Editing Programs and the solution of the Data Quality Problems. Stockholm: Sweden.

Granquist, L. (1984). On the Role of Editing. Statistik Tidskrift, 2, 105-118.

Granquist, L. (1990). A Review of some Macro-Editing Methods for Rationalising the Editing Process. Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality Oct 1990, 225-234.

Granquist, L. (1994). Macro-Editing - A Review of Methods for Rationalising the Editing of Survey Data. Statistical Data Editing Volume One, Methods and Techniques. (1994) New York: United Nations, 111-126.

Granquist, L. (1995). Improving the Traditional Editing Process. Business Survey Methods, Editors Cox et al. New York: John Wiley and Sons.

Granquist, L. (1997). The New View on Editing. International Statistical Review. Revue International de Statistique. Vol 65, No 3, 381-387

Granquist and Kovar, (1997). Editing of Survey data: How much is Enough? Survey Measurement and Process Quality, L Lyberg et al, New York: Wiley, 415-171.

Greenberg, B. (1986) The Use of Implied Edits and Set Covering in Automated Data Editing. Washington: US Bureau of the Census.

Hunter, Lecily and Ladds, John. (1995). Editing Survey Data After CAI Collection. Ottawa: Statistics Canada.

Kish, L. (1967). Survey Sampling. New York: John Wiley and Sons.

Kovar, John and Winkler, William (1996) Editing Economic Data. American Statistical Association, 1996 Proceedings of the Section on Survey Research Methods, Alexandria: American Statistical Association, 81-89.

Lepp, H and Linacre, S. (1993). Improving the Efficiency and Effectiveness of Editing in a Statistical Agency. Bulletin of the International Statistical Institute: Proceedings of the 49th Session, Florence, Italy Book 2, 111-112.

Rowland, John. (1994, updated 1998). Labour Force Survey Redesign. The Strategy for the LFS Editing and Imputation. Ottawa: Statistics Canada.

Statistics Canada Quality Guidelines, Third edition. (1998). Ottawa: Statistics Canada.

United States Federal Committee on Statistical Methodology. (1990). Paper 18: Data Editing in Federal Statistical Agencies. Web site: Federal Committee on Statistical Methodology.

van de Pol, Frank and Bethlehem, Jelke. (1997). Data Editing Perspectives. Statistical Journal of the United Nations Economic Commission for Europe, Vol 14, No.2, June 1997, 153-171.

Whitridge, P and Kovar J. (1990). Applications of the Generalized Edit and Imputation System at Statistics Canada. 1990 Proceedings of the Section on Survey Research Methods, American Statistical Association, 105-110.

Williams, Howard. (1999). Survey of Income and Housing Costs. Review of Editing and Imputation. Report and Recommendations. Australian Bureau of Statistics, unpublished document.